International Journal of Computer Science Engineering
and Information Technology Research (IJCSEITR)
ISSN (P): 2249-6831; ISSN (E): 2249-7943
Vol. 7, Issue 5, Oct 2017, 9-28

HYBRID KERNEL FUZZY CLUSTERING WITH FEED LION NEURAL NETWORK FOR MISSING DATA IMPUTATION AND CLASSIFICATION

R RAJANI[1] & T.SUDHA[2]

[1]*Department of MCA, Narayana Engineering College, Nellore, India*

[2]*Department of MCA, Sri Padmavathi Mahila University, Tirupati, India*

ABSTRACT

A common issue in many practical applications associated with pattern classification is data incomplete or missing data due to various reasons that differ based on the applications. Missing data imputation is a promising approach used to handle this issue, which fills the missing attributes with estimated values following several techniques. This paper proposes a new strategy for data imputation and classification using a hybrid prediction model that combines Hybrid Kernel Fuzzy Clustering (HKFC) and Feed Lion Neural Network (FLNN). FLNN considers the missing attributes as the class attributes to predict the missing data, which will also be computed using HKFC and finally, the combined effect of these two approaches provides the missing data. FLNN is designed for the classification by modifying Leven berg-Marquardt (LM) - based feed forward neural network with the incorporation of LOA and thus, selects the weights optimally. The experiment is carried out using three data sets from UCI, based on two metrics, Mean Squared Error (MSE) and accuracy. The performance of the proposed method, HKFC+FLNN is compared with three techniques, FC+FLNN, FC+KNN and K-Means+KNN, where HKFC+FLNN attained MSE by 0.0344, 0.0048, 0.2754, and classification accuracy of 0.8022, 0.95 and 0.9741, for heart disease, iris, and wine datasets, respectively.

KEYWORDS: Data Incompleteness, Missing Data Imputation, FCM, Classification & Neural Network

Received: Jul 28, 2017**; Accepted:** Aug 18, 2017**; Published:** Sep 29, 2017**; Paper Id.:** IJCSEITROCT20172

Original Article

INTRODUCTION

Since the data mining techniques are non-trivial approaches that discover new data patterns and relationships, they are used commonly for gaining knowledge in databases [7]. One of the issues in data mining is data quality, that exists in all types of databases as data incompleteness or missing data [1, 9]. Missing data imputation is a process that provides estimations for the missing data, from the data observed by analyzing the data [20]. As missing data may lead to bias that affects the classification performance and the quality of the learned data patterns, missing data imputation has become a major concern in learning incomplete data [11, 25]. The missing data approaches are of three types: Missing At Random (MAR), missing completely At Random (MCAR) and Not Missing At Random (NMAR). In MAR, the probability of the missing data is based on other observed responses, whereas, MCAR does not relate the Missing Values (MVs) with other responses observed in the dataset. Missing data in NMAR takes place only on certain values of interest and so, it is based on those values itself [16]. The missing values can be dealt in a simple manner only when i) the data has a small number of observations that contain MVs, or if ii) the analysis does not result in a serious bias [10].

Data imputation refers to the task of substituting the MVs in the dataset with the aim of reducing the bias of survey estimates. Data imputation techniques lessen bias and, thereby, the standard analysis can be done using a

dataset [13]. The missing data issues can be solved using several approaches as follows: i) Reject the objects that contain missing data, ii) Adopt a manual approach to fill the gaps, iii) Replace the values with a constant, iv) Take the mean or the mode of the objects for the replacement, and v) Fill the missing data with the most probable value [12]. One of the main characteristics in data imputation is the MV pattern, as it decides the selection of the imputation method. Data imputation can be classified into two kinds, such as single imputation and multiple imputation [14]. Single imputation methods, like mean imputation, hot-deck imputation and regression, are those that denote the replacement of a missing value with a single value. In multiple imputation, each missing data are substituted with a set of possible values taken from observed data, providing multiple data sets that cause the imputation uncertainty to be integrated into statistical inferences [15].

Nowadays, several machine learning imputation techniques have been designed. The classification based on machine learning techniques is an important process in data mining and pattern recognition. Various solutions to the classification have been created using different classifiers for different applications, such as monitoring, web log analysis, traffic management, telecommunication, medical data analysis, etc. [5]. Some of the machine learning approaches is K-Nearest Neighbor (KNN) [2], Self-Organizing feature Map (SOM) [5], multi-layer perceptron [17], auto-associative neural network imputation with genetic algorithms [18] and fuzzy-neural network. They build a predictive model to evaluate the MVs by estimating the values that are absent based on the information in the dataset [4]. If the missing attribute in the missing data imputation is nominal, the learning process becomes classification, whereas it is a regression, if the attribute is continuous. A machine learning technique is trained for every instance with a missing attribute depending on the instances without MVs and the model uses non-missing values of the instance to predict the missing attribute [19]. However, it has two main restrictions as follows: i) imputation methods other than the modelling task cannot be estimated properly, and ii) Removal of an instance or attribute information with MVs, called complete-case analysis, must be avoided [22].

This paper intends to design a missing data imputation and classification methodology using a hybrid model that combines HKFC and FLNN. This approach involves two phases: i) Clustering and neural network combined to miss data imputation and ii) Hybrid neural network based classification. Based on the cluster centroids estimated by HKFC, the data records can be clustered for the prediction of missing data. In the feed-forward neural network that uses a LM training algorithm, LOA is incorporated to improve the performance of data imputation and classification. By the effective combination of those predicted data, the missing data is imputed. Once the missing values are obtained, the proposed FLNN approach classifies the data by selecting the appropriate weights using the integrated LOA.

The main contributions of the proposed missing data imputation and classification technique are as follows:

- Imputation of missing data using two combined strategies: HKFC, for an effective feature clustering and feed forward neural network that uses LM and LOA, for the computation of missing data in the data record.
- Introduction of the hybrid network model, FLNN, by the utilization of an optimization algorithm, LOA, to assign optimal weights in the neural network, for the accurate classification of desired data based on its application.

The rest of the paper is structured as given: Section 2 presents the literature survey, where the existing missing data imputation and classification techniques are reviewed. Section 3 explains the proposed MKFC and FLNN- based approach for computing missing data and classification approach with a suitable block diagram. Section 4 discusses the results of the proposed method with performance comparisons, and section 5 concludes the paper.

MOTIVATION

Related Works

This section deals with the techniques and algorithms that were used in the existing systems for the data imputation to get an idea for the implementation of the proposed data imputation approach.

Zhixu Li *et al.* [1] presented a technique, called in Teractive Retrieving-Inferring data imputation (TRIP) by analyzing the interaction between the approaches based on inferring and retrieving to fill the missing attributes in the dataset. The imputation recall was enhanced even when a small number of selected MVs were retrieved. This technique could increase the imputation recall of inferring-based schemes. However, inferring of un-inferable MV is impossible even if all the other MVs are known.

Jose Luis Sancho-Gomez *et al.* [2] Developed a KNN imputation method using a metric that utilized feature-weight, distance based on Mutual Information (MI). This technique provided a measure based on missing data for the classification using an imputed dataset to increase the performance. Even though the efficiency can be enhanced in both artificial and real classification datasets, the searching process is difficult when the KNN classifier imputes for the similar instances.

Chandan Gautam and Vadlamani Ravi [3] designed two hybrid imputation techniques using Particle Swarm Optimization (PSO), Auto Associative Extreme Learning Machine (AAELM) and Evolving Clustering Method (ECM) that also preserved the covariance structure of the data. By adopting ECM between the input and the hidden layer, randomness of AAELM was removed. The imputation performance was enhanced using the activation function. The major drawback is that it failed to maintain the covariance structure of the original data.

Loris Nannia *et al.* [4] had presented a multiple imputation approach depending on random subspace, where the missing data were computed using a different data cluster. The clustering method adopted was fuzzy clustering. The performance of this approach was evaluated by comparing with various statistical and machine learning imputation techniques using different datasets. The limitation of this approach is that the observation was not considered in the situation where the number of features and samples were larger.

Laura Folguera *et al.* [5] Developed a Self-Organizing Map (SOM) based technique of data imputation based on the distance object per one weight to estimate physicochemical metrics of water samples in a dataset that had concentrations of various analysts that were missed. The technique was based on two possibilities: i) considering sample vectors in the training data set with and without missing data and ii) pre-training a SOM for a dataset without MVs and then, perform imputations for another sample data set with MVs. Different data variables that were missing could be imputed at the same time, but, large proportions of MVs might result in erroneous outputs.

Jaemun Sim *et al.* [6] Presented an adaptive technique to select a suitable classification algorithm/imputation approach pair to recognize the dataset features and to make changes automatically when needed. The technique was designed by modifying case-based reasoning in the following ways: the original case base was pre-processed to construct a null data structure. Then, multiple pairs were identified to form a candidate set, from which a pair was selected. Even though the technique is scalable and accurate, it consumes considerable time when the number of matching is increased.

Gerhard Tutz and Shahla Ramzan [7] adopted a nonparametric approach, NN imputation, modifying it based on a

distance measure. It had only a negligible imputation error than other NN approaches and considered weighted NN imputation techniques that used distances for selected covariates. The performance can be improved with an appropriately selected distance that carries information about MVs. However, the lower correlation may degrade the performance.

Md. Geaur Rahman and Md Zahidul Islam [8] designed a technique, Fuzzy Expectation Maximization and Fuzzy Clustering-based Missing Value Imputation Framework for Data Pre-processing (FEMI). FEMI imputed numerical and categorical MVs depending on the records that had MVs based on a guess. The quality of imputation is higher, but the process is difficult when the nature of MVs is random.

Problem Statement

Let D be the database that contains a number of data records, as $D = \{D_1, D_2, \ldots, D_k\}$, such that $1 \le i \le k$, where k is the total number of records. Each record in the database has n attributes, expressed as $D_i = \{b_{i1}, b_{i2}, \ldots, b_{in}\}; 1 \le j \le n$. By clustering the input database into s clusters based on the centroids and then, by averaging, the missing attribute values can be predicted. Similarly, a neural network computes the missing data, which is combined with that in the clustering approach for data imputation. Once the missing data are identified, the FLNN-based classifier performs the classification.

Challenges

The problems associated with data acquisition, sensor network problems, measurement errors, data migration failures, and so on are the major reasons for the cause of the most diverse effect, called missing data [22]. Some of the challenges noticed from the above literature review are stated as follows:

- One of the challenges in the field of data analysis is the presence of missing data [15]. The objective of any data imputation approach is to fill the missing values in a data set such that a standard data analysis technique applied can analyze the completed data set.
- The second challenge in data mining is the missing data imputation technique itself, as the presence of MVs affects the data quality, generate bias and so on [18, 21].
- MVs in data classification can result in various serious issues during the learning process, like inefficient classification, biased data structure, difficulty in the analysis and degeneration of prediction performance [19].
- Missing data in the training set and/or in the test set lessen the prediction accuracy of the learned classifiers, as it depends on the proportion of missing data and the high dimensionality [23, 24].
- Designing an efficient classifier that can handle the datasets, which changes the patterns of the missing data, volume and structure of the data constants, is another challenge [6].

PROPOSED CLUSTERING AND HYBRID NETWORK MODEL-BASED MISSING DATA IMPUTATION AND CLASSIFICATION APPROACH

The proposed technique of imputing missing value and classification is explained in this section in two phases. In the first phase, missing data imputation is carried out by combining a clustering approach, HKFC, with the neural network. The missing attributes in the input data are considered as the class attributes for the computation of MVs. Secondly, a

hybrid network approach, FLNN, is employed to select the weights of neurons in the network for the classification. Thus, the new training algorithm can effectively perform the classification by predicting the missing data accurately. Figure 1 shows the block diagram of the proposed data imputation and classification approach that utilizes clustering and hybrid network model. The further sections elaborate the two main contributions of the proposed technique.

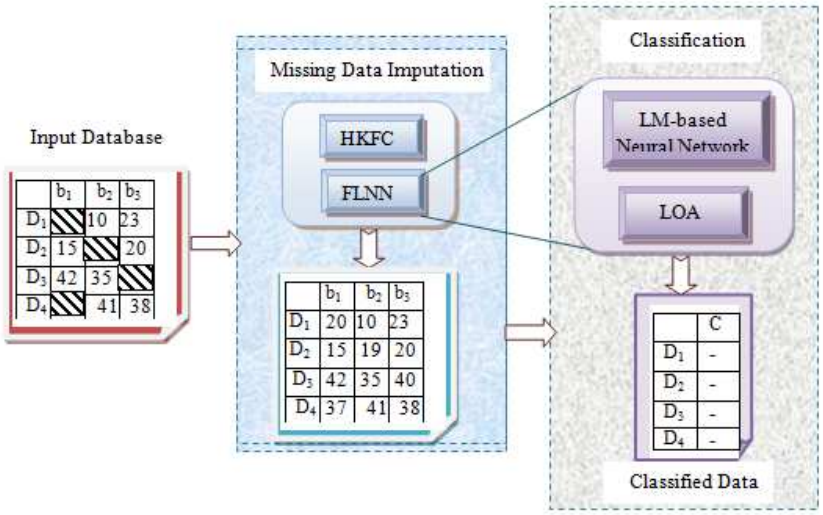

Figure 1: Block Diagram of Proposed HKFC and FLNN Based Missing Data Imputation and Classification Technique

Missing Data Imputation by the Combined Effect of HKFC and Neural Network Model

This part deliberates the first contribution of the research, i.e. missing data imputation, which attempts to solve the common data quality issue, called data incompleteness. The techniques used for this purpose are a clustering approach, HKFC and a feed forward neural network. MKFC extends Fuzzy C- Means (FCM) [28] by adopting multiple kernel functions in its membership degree for clustering. By utilizing the averaging process of centroids, MKFC can compute the missing attributes, which are then, combined with the values generated by the neural network, for the effective data imputation. These two techniques that are utilized for the data imputation are described in the sections below.

HKFC-Based Missing Data Prediction

One of the flexible and commonly employed clustering approaches is FCM. However, it is not much applicable for general clusters other than spherical. Hence, kernel-based techniques are adopted in FCM, as an extension, that can be used for higher dimensional feature space. MKFC [26] is an improved clustering approach that avoids the confusion in choosing suitable kernels, which happens in kernel integrated FCM clustering. Based on the procedure of MKFC, HKFC is developed that could find the most appropriate degrees of membership and thereby, enhances the clustering process. An objective function that is to be minimized is defined based on the membership degree and the distance as follows,

$$F_r = \sum_{j=1}^{M} \sum_{m=1}^{N^C} v_{jm}^r \left\| y_j - c_m^l \right\|^2 \quad (1)$$

where, v_{jm} is the degree of membership, y_j represents the j^{th} u-dimensional vector in the data set Y, c_m^l is the center of m^{th} cluster, N^C is the total number of clusters and M is the total number of data in Y.

Let $V = \left[v_{jm}\right]$ be the membership matrix, such that the matrix the limit secified within limits, $1 \le j \le M; 1 \le m \le N^C$. The solution to the optimization problem is brought here by updating the matrix V as in equation (2),

$$v_{jm} = \frac{1}{\sum_{p=1}^{N^C} \left[\frac{d\left(y_j, c_m^l\right)}{d\left(y_j, c_p^l\right)} \right]^{\frac{2}{r-1}}} \tag{2}$$

Where, $d(.)$ is the distance measure between any data with its center and r is a real number that takes a value larger than 1, i.e. $r > \infty$. In HKFC, the distance measurement utilizes the kernel function that makes it applicable in many applications due to its ability in handling unreliable features. This is represented by an exponential and the tangent function, as in equation (3),

$$d\left(y_j, c_m^l\right) = \left[\exp\left\|y_j - c_m^l\right\| + \tanh\left\|y_j - c_m^l\right\|\right] \tag{3}$$

Then, the cluster center is defined by a ratio regarding the membership degree and the data as follows,

$$c_m^l = \frac{\sum_{j=1}^{M} v_{jm}^r y_j}{\sum_{j=1}^{M} v_{jm}^r} \tag{4}$$

The update process is repeated until it reaches a termination condition, which is given here based on the difference between the membership matrices obtained at iterations t and $t+1$. The procedure will terminate when the difference is less than δ, where δ is a constant within the range $[0,1]$. Based on the clustering performed, by the averaging process of centroids, the missing data can be obtained as expressed below,

$$b_{ij}^* = \frac{1}{n_C} \sum_{l=1}^{n_C} b_{lj} \tag{5}$$

Where, n_C is the number of data points belonging to the cluster of i^{th} data?

Predicting Missing Attributes using Neural Network Model

With the missing attributes as the class attributes, a neural network with incorporated LOA will be trained for the prediction of missing data. Consider a feed forward network with LM-based training algorithm, shown in figure 2, where the weight update is modified using the LOA. In training, the features used learn the neurons by selecting optimal weights using the adopted optimization algorithm and the class labels are found based on the learned weights. The network is comprised of three layers: input, hidden, output, where the number of hidden layers is considered to be one. The missing attribute prediction procedure using FLNN is explained as follows.

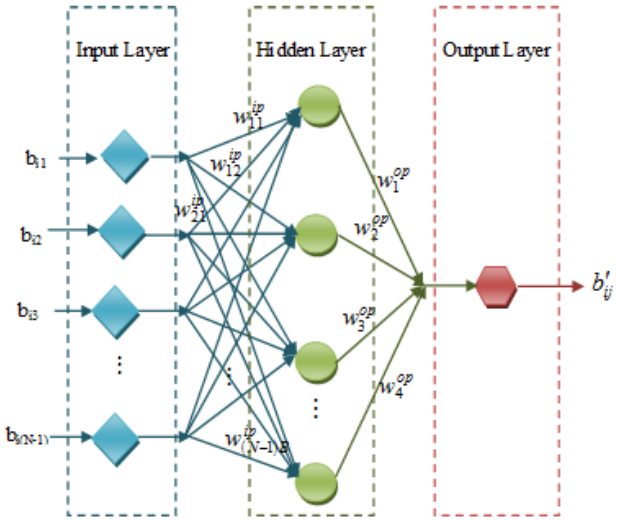

Figure 2: Adaptive Neural Network Architecture

Let $b = \{b_{i1}, b_{i2}, \ldots, b_{i(N-1)}\}$ be the input features fed to the neural network and b_{ij}^{o} be the output vector, which is the missing data to be identified. For B bias weights in the hidden layer, let the weight vector w can be represented within the range $(N-1)\times B + B\times 1$ as,

$$w = \{w_q ; 1 \le q \le (N-1)\times B + B\times 1\} \tag{6}$$

Based on the optimally selected weights in FLNN, the output vector can be computed as follows,

$$b'_{ij} = \sum_{j=1}^{N} [g_j * w_j^{op}] \tag{7}$$

Where, g_j is the j^{th} neuron in the hidden layer, is the total number of neurons and w_j^{op} is the weight function at the output. The hidden layer neurons depend on the input features and the weights that connect the input layer to the hidden layer, which can be represented as,

$$g_j = \sum_{j=1}^{N} [b_{ij} * w_j^{ip}] \tag{8}$$

Where, b_{ij} is the input feature and w_j^{ip} is the input weight connecting input vectors with the hidden layers.

HKFC-FLNN Based Missing Data Imputation

For the imputation of missing data, the data predicted by both the approaches discussed in sections 3.1.1 and 3.1.2 are combined, based on two weight coefficients. By the averaging process, HKFC could predict the missing attributes, whereas, FLNN predicted the values based on weight assignment. Finally, by averaging the data predicted by both the techniques, the missing data is computed. This can be represented mathematically as,

$$b_{ij}^{o} = \frac{\alpha b_{ij}^{*} + \beta b_{ij}'}{\alpha + \beta} \tag{9}$$

Where, b_{ij}^{*} is the HKFC-based predicted missing data, b_{ij}' is the missing data predicted using FLNN, α and β are the weighted coefficients ranging between 0 and 1.

This process can be explained by considering a running example as given below,

Running Example: The proposed technique of computing missing data is illustrated in figure 3.

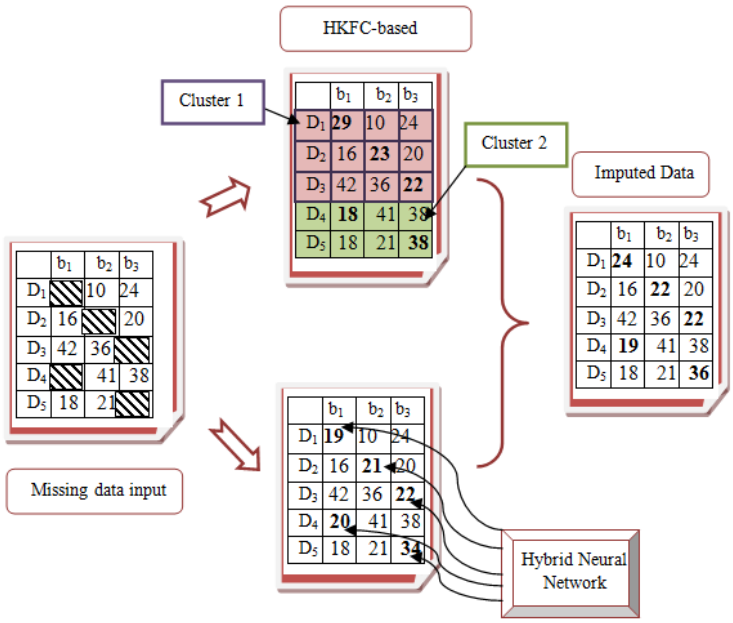

Figure 3: Missing Data Imputation

As shown in figure 3, in HKFC approach, the data set is divided into two clusters, i.e., D_2 and D_3 are grouped into a cluster (cluster 1) and D_4 - D_5 form another cluster (cluster 2). The missing data in b_1 of cluster 1 is computed by averaging the data obtained for other attributes in cluster 1 of b_1. For example, the data obtained for the attribute b_1 is $D_2 = 16$ and $D_3 = 42$. Hence, the missing data D_1 can be computed by averaging both the values, so that $D_1 = 29$. Similarly, for cluster 2, the same procedure is repeated. Thus, HKFC predicts the missing data for all the attributes in the dataset. The same input dataset is fed to FLNN, which predicts the data based on the weights selected.

Proposed LONN Model Based Classification

In this section, the proposed neural network model, FLNN, for the classification is explained in detail. Once the process of missing data imputation is completed, the classification of data is performed. One of the popular artificial intelligence models [30] used for the classification is a feed forward neural network [29]. The proposed classification approach utilizes an adaptive neural network that uses a LM training algorithm, where the neuron weight assignment is based on the update rule of LOA. LOA is a meta - heuristic approach inspired from the social organization and behavior of lions. The algorithm is adopted here, for the selection of feasible weights based on the error estimate. The classification

procedure using FLNN is illustrated as given in the following steps.

Initialization

The fundamental step in modelling FLNN is the initialization of network architecture. Let $b = \{b_1, b_2, \ldots, b_N\}$ be the feature a number that provides N number of features as inputs to the network. Let $z = \{z_1, z_2, \ldots, z_M\}$ be the output vector, where M represents the dimension of the output vector, which is the total number of classes required. The weight vector is then initialized in random as given in equation (6).

Error Estimation

The error measure determines the accuracy of classification by allowing the neural network to choose the proper weight coefficient using the algorithm. In LM-based neural network, the error estimation is given regarding the difference between the output vector and the ground truth value.

$$E(t) = \frac{1}{N}\sum_{i=1}^{N}\left[z_i - z_i^G\right]^2 \tag{13}$$

where, z_i is the output function, which is represented as z_i^{LM} for LM-based neural network and z_i^G is the ground truth value. The error estimate using LM approach at instant t is denoted as $E^{LM}(t)$. The output vector z_i^{LM} is given in the form a function that depends on the input feature vector and the weight vector as in equation (14),

$$z^{LM} = F\left(b, w^{LM}\right) \tag{14}$$

Where, b is the feature vector given as input to the network and w^{LM} is the LM-based weight vector. In FLNN, the update rule of LOA is utilized, which also requires error estimation, similar to that in equation (13), as formulated below,

$$E^{LOA}(t) = \frac{1}{N}\sum_{i=1}^{N}\left[z_i^{LOA} - z_i^G\right]^2 \tag{15}$$

Where, z_i^{LOA} is the output obtained when w^{LOA} is the weight applied and it is given as,

$$z_i^{LOA} = F\left(b, w^{LOA}\right) \tag{16}$$

Where, w^{LOA} is the weight computed at iteration t by adopting LOA.

The weight update in FLNN will be determined by the error estimates obtained using equations (13) and (15). If $E^{LM}(t) > E^{LOA}(t)$, then the weight update is done using w^{LOA}, which is to be derived later.

Weight Update using LM Algorithm

The important process in neural network model is the neuron weight assignment, which must be carried out using an effective approach for an accurate classification. Once the network is initialized, the weights are updated by following LM training approach, in every iteration, as given below,

$$w^{LM}(t+1) = w(t) - [S + \mu * T]^{-1} * H \tag{10}$$

Where, $w(t)$ is the weight computed at iteration t, S is the Hessian matrix, μ is the damping factor, T is the identity matrix and H is the gradient. The hessian and the gradient matrices are given based on Jacobian matrix, as follows,

$$S = G^T * G \tag{11}$$

$$H = G^T * E \tag{12}$$

Where, G is the Jacobian matrix and E is the error estimate.

Weight Update using LOA

To improve the search process and the accuracy, LOA [27] is employed in the update rule, where it selects the most suitable weights for the network neurons. Following the hunting behavior of lion in LOA, the weight update at iteration $t+1$, for FLNN, is derived as,

$$w^{LOA}(t+1) = w(t) + e \times P \times [w(t) - w^{LM}(t)] \tag{17}$$

Where, e is a number selected at random from the range $[0,1]$, $w^{LM}(t)$ is the weight vector estimated at previous iteration using LM algorithm. P denotes the percentage of improvement, which is defined regarding the ratio of error estimates, as follows,

$$P = \left[\frac{E(t) - E(t-1)}{E(t)} \right] \tag{18}$$

Where, $E(t)$ and $E(t-1)$ are the error estimates computed at time t and $t-1$, respectively.

Termination

The process is repeated for a maximum number of iterations or when no significant solutions can be obtained. The weights are selected in FLNN in such a way that the error computed at current iteration, is reduced from that calculated at the previous iteration. If $E(t) < E(t-1)$, then the damping factor μ is reduced by a value f and vice-versa.

The pseudo code of the proposed FLNN approach for the classification is demonstrated in table 1.

Table 1: Pseudo code of Proposed FLNN Model

	FLNN Algorithm
1	Input: Feature vector b and ground truth value z^G
2	Output: Weight coefficient w
3	Begin
4	Initialize the weight vector
5	for $(t < \max\_iteration)$
6	Calculate the weights using equations (10) and (17)
7	Compute the errors E^{LM} and E^{LOA} using equation (13)

8	if $\left(E^{LOA} < E^{LM}\right)$
9	$w(t+1) = w^{LOA}(t+1)$
10	else
11	$w(t+1) = w^{LM}(t+1)$
12	end if
13	if $\left(E(t) < E(t-1)\right)$
14	$\mu = \mu - f$
15	else
16	$\mu = \mu + f$
17	end if
18	$t = t+1$
19	end for
20	Return $w(t+1)$
21	Terminate

RESULTS AND DISCUSSIONS

This section demonstrates the results of the proposed approach of data imputation and classification using a comparative analysis to estimate the performance level of the techniques in imputing and classifying the missing data.

Experimental Setup

The experimentation of the proposed HKFC-FLNN approach is done in a system of Windows 8 OS with the configurations of Intel i-3 core processor and 2 GB RAM. The algorithm is implemented using MATLAB and the performance is evaluated using the metrics, MSE and accuracy.

Dataset Description

The experiment is executed using three data sets, Heart disease (Dataset 1), Iris (Dataset 2) and Wine (Dataset 3), taken from the UCI machine learning repository. *Dataset 1:* This dataset, donated by David W. Aha, has 14 commonly used attributes from a total of 76 and four databases, Cleveland, Hungary, Switzerland, and VA Long Beach, among which, Cleveland is the one used widely. The dataset points out the presence of heart disease using an integer given from 0, which indicates the absence of heart disease, to 4.

Dataset 2: The iris dataset is created by R.A. Fisher and donated by Michael Marshall. It contains 150 instances and 4 attributes with Fisher's paper as the frequent database used till date.

Dataset 3: The third dataset obtained from the chemical analysis result of wines consists of 178 instances and 13 attributes, such as alcohol, malic acid, ash, alcalinity of ash, magnesium, total phenols, flavanoids, nonflavanoid phenols, proanthocyanins, color intensity, hue, OD280/OD315 of diluted wines and proline. The description of these three datasets is given in https://archive.ics.uci.edu/ml/datasets.html

Evaluation Metrics

The parameters used for the evaluation of performance of classification and prediction are two, MSE and accuracy, which are defined as follows,

MSE: It evaluates the average of the squares of the difference between the estimated output and that to be

estimated and is represented as,

$$E(t) = \frac{1}{N}\sum_{i=1}^{N}\left[z_i - z_i^G\right]^2$$

Accuracy: The classification accuracy can be measured in a ratio given by,

$$Accuracy = \frac{T_P + T_N}{T_P + F_P + T_N + F_N}$$

where, T_P is True positive, which is recognized correctly, T_N is True negative, which is correctly rejected, F_P is False positive, which is recognized incorrectly and F_N is False negative, which is correctly rejected.

Methods Employed for Comparison

The performance of the proposed method is compared with that of three existing techniques, FC+FLNN, FC+KNN and K-Means+KNN. In [8], Fuzzy Clustering (FC) is used for finding missing data, whereas in [2], KNN could perform both the classification and imputation of missing data simultaneously. K-Means [23] performs missing data imputation based on a weighted distance.

Experimental Results

The results of imputing and classifying missing data are given regarding a comparative analysis, where the experiment carried out in three datasets is presented. The analysis is performed using the two evaluation metrics based on varying percentage of missing data and cluster size.

Analysis Based on the Percentage of Missing Data

The comparative analysis based on the percentage of missing data is made in the three datasets using MSE and accuracy, as explained below,

MSE

In this part, the analysis made on all the techniques used for evaluation in the three datasets regarding MSE is given. The technique having minimum MSE indicates the effectiveness of the approach. Figure 4 shows the results of the analysis in plots of MSE against the percentage of missing data in HKFC+FLNN and in three existing techniques, FC+FLNN, FC+KNN and K-Means+KNN. In figure 4.a, the comparative analysis plot of MSE for dataset 1 is given against the percentage varying from 10 to 50. Initially, when the percentage of missing data is 10, the error value in FC+FLNN, FC+KNN and K-Means+KNN is 0.1558, 0.1477 and 0.0727, whereas, in HKFC+FLNN, it is just 0.04. When the percentage is kept to its maximum, MSE obtained in FC+KNN and K-Means+KNN is increased to 0.1496 and 0.0803. Meanwhile, in the proposed HKFC+FLNN approach, it is reduced to 0.037. In the MSE analytical curve of dataset 2 pictured out in figure 4.b, K-Means+KNN had the highest error of all against 10% and 50% missing data, with values 0.0244 and 0.0242, which reaches 0.0318 when the missing data percentage is 40. HKFC+FLNN, FC+FLNN and FC+KNN had MSE values 0.0075, 0.0121 and 0.0104 initially, which rises to 0.0088, 0.0154 and 0.0134, respectively, for the maximum percentage. Figure 4.c presents the analysis based on MSE for the third dataset. The MSE value of HKFC+FLNN had reduced from 0.3139 to 0.2886 when the percentage of missing data are raised from 10 to 50.

Meanwhile, the existing FC+FLNN, FC+KNN and K-Means+KNN had MSE values of 0.3656, 0.3256 and 0.3456, which increased to 0.4107, 0.3889 and 0.4317.

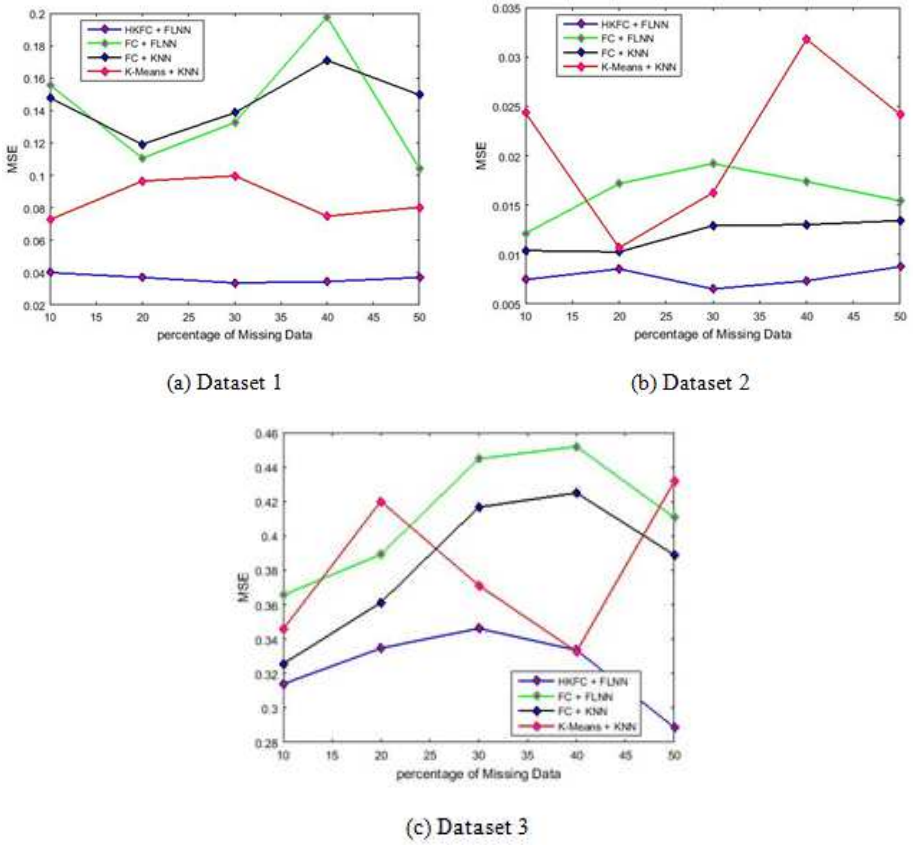

Figure 4: MSE Vs Percentage of Missing Data

Classification Accuracy

Figure 5 presents the classification accuracy calculated for all the comparative techniques in the three datasets. In figure 5.a, the accuracy analysis plot against different percentage of missing data for the heart disease dataset is shown. When the maximum classification accuracy of FC+FLNN, FC+KNN and K-Means+KNN is 0.7429, 0.7319 and 0.7582, the proposed HKFC+FLNN approach had attained an accuracy of 0.8022. The classification accuracy result obtained in all the approaches for dataset 2 is depicted in figure 5.b. As shown in the figure, the accuracy of all the three existing techniques is at the same point 0.9111 despite of varying percentage of missing data. However, HKFC+FLNN could attain a value 0.95, which is 0.0409% more than the accuracy of the existing techniques. Figure 5.c illustrates the accuracy analysis plot for the wine dataset in the four comparative techniques. The maximum accuracy obtained in HKFC+FLNN, FC+FLNN, FC+KNN and K-Means+KNN is 0.9244, 0.8741, 0.8852 and 0.9074. This suggests that the proposed technique has the highest accuracy when compared with the existing schemes and thus, proves its effectiveness.

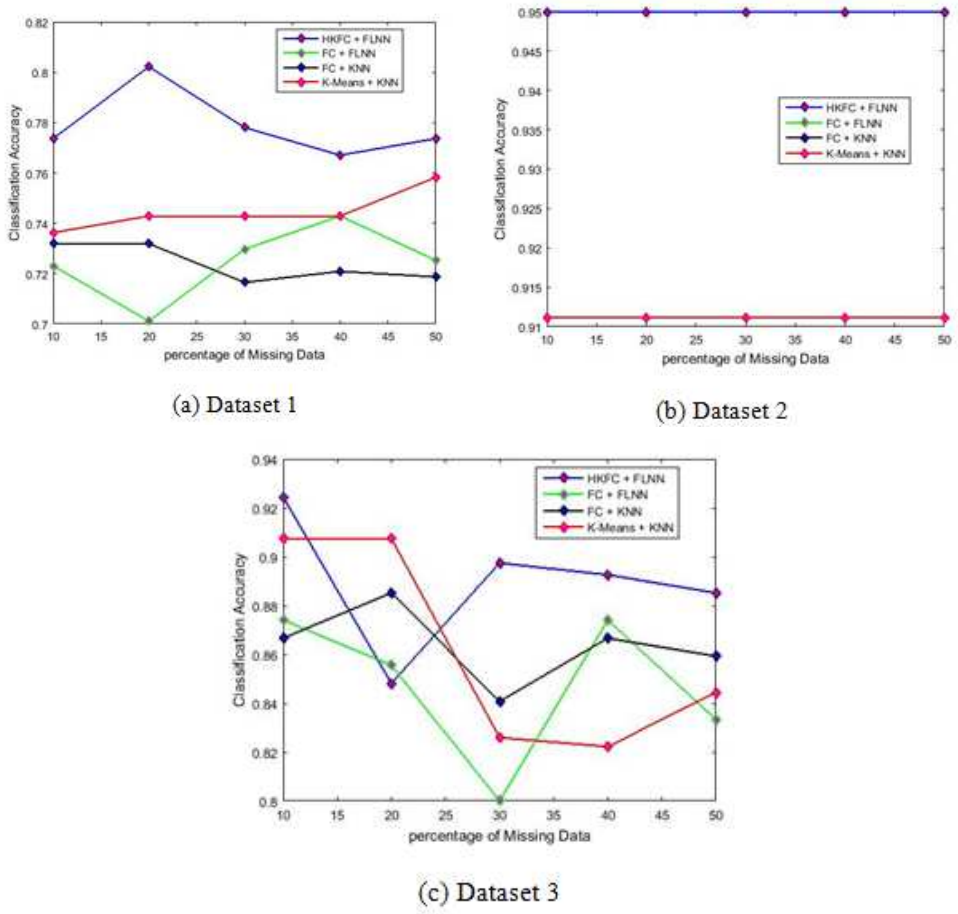

Figure 5: Classification Accuracy Vs Percentage of Missing Data

Analysis Based on the Cluster Size

Another comparison of performance is made by analyzing the evaluation metrics based on the cluster size varied from 1 to 5. The results obtained are discussed in the further subsections as follows,

MSE

The comparative analysis based on MSE on the four techniques for heart disease, iris and wine datasets, is demonstrated in figure 6. Figure 6.a shows the error analysis graph for dataset 1 in the techniques, HKFC+FLNN, FC+FLNN, FC+KNN and K-Means+KNN. When the cluster size is fixed to 1, FC+FLNN has shown a maximum MSE of 0.1179, while, the proposed technique has just 0.0437. As the size of the cluster is 5, it is reduced to 0.0357 in HKFC+FLNN, whereas in FC+FLNN, the error is 0.1086. In figure 6.b, the analysis result based on MSE is presented, where the minimum MSE value obtained in the existing FC+FLNN, FC+KNN and K-Means+KNN strategies are 0.0057, 0.0101 and 0.0061, while HKFC+FLNN has a value 0.0048 as the minimum MSE. The comparative analysis chart based on MSE for the wine dataset is given in figure 6.c. Initially, for the cluster size 1, FC+FLNN, FC+KNN and K-Means+KNN have MSE of 0.3489, 0.3208 and 0.3725, whereas, HKFC+FLNN has 0.2754. The error value is found to be increasing in HKFC+FLNN, FC+FLNN and FC+KNN, as the cluster size is kept 5, with 0.308, 0.3913 and 0.3521, respectively. However, the proposed HKFC+FLNN method is proven to be better, as it shows the least error compared to that of existing approaches.

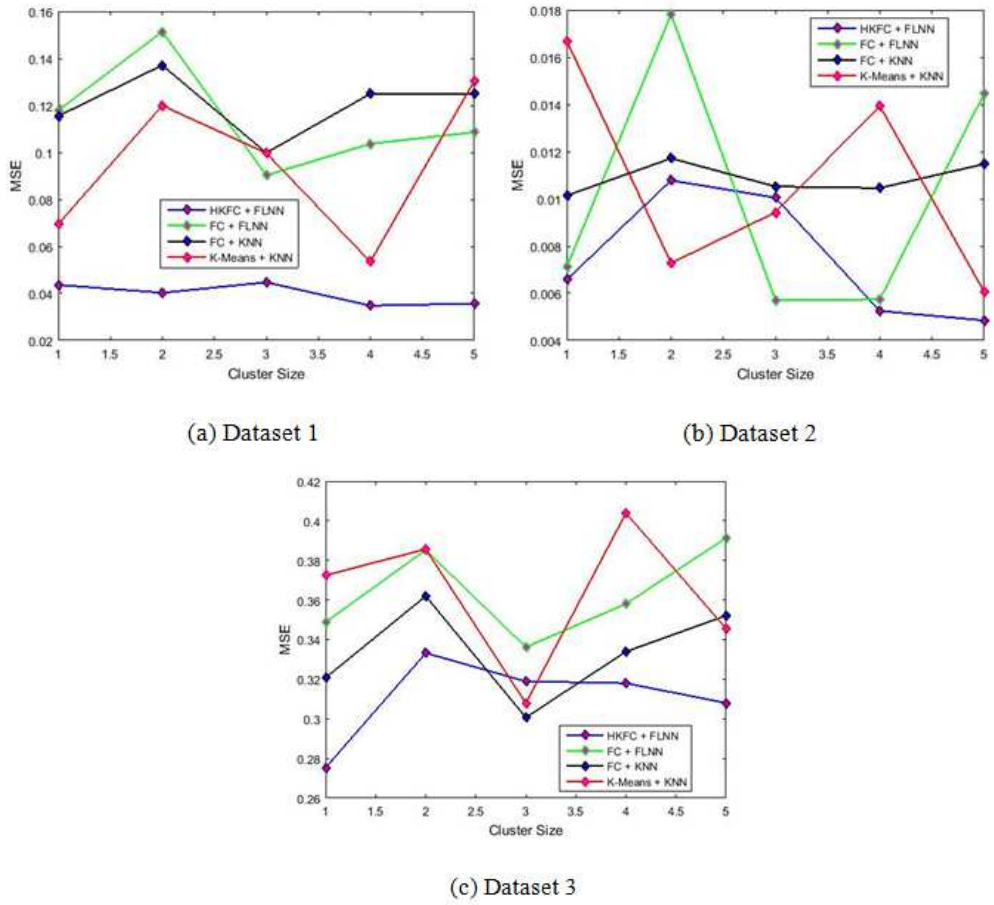

Figure 6: MSE Vs Cluster Size

Classification Accuracy

The analysis results of accuracy for the three datasets by varying the cluster size in the comparative methodologies are pictured out in figure 7. Figure 7.a shows the classification accuracy result for the dataset 1, where HKFC+FLNN, FC+FLNN, FC+KNN and K-Means+KNN have shown an accuracy of 0.7846, 0.7385, 0.7297 and 0.7516, respectively, for the cluster size 1. When the cluster size is 5, HKFC+FLNN provides an accuracy of 0.7758, which is 0.0367% more than the accuracy of FC+FLNN that had a value of 0.7473. In figure 7.b, the accuracy analysis graph for the second dataset is given. Similar to the first case, i.e. in the analysis based on varied missing data percentage, the accuracy Vs cluster size shows the same result with overlapped values in the existing techniques. The proposed HKFC+FLNN approach has an accuracy of 0.95, while the remaining schemes have attained 0.9111. The classification accuracy result for the third dataset is depicted in figure 7.c. Here, for the cluster size 1, the existing methods, FC+FLNN, FC+KNN and K-Means+KNN had an accuracy of 0.8259, 0.8481 and 0.8741, respectively, while HKFC+FLNN had 0.8852. As the cluster size is varied, the proposed HKFC+FLNN could offer a maximum accuracy of 0.9741, while FC+FLNN, FC+KNN and K-Means+KNN could produce only 0.8741, 0.9111 and 0.8815. Hence, it is clear that the proposed approach has maximum classification accuracy and thus, a better performance than the other methods.

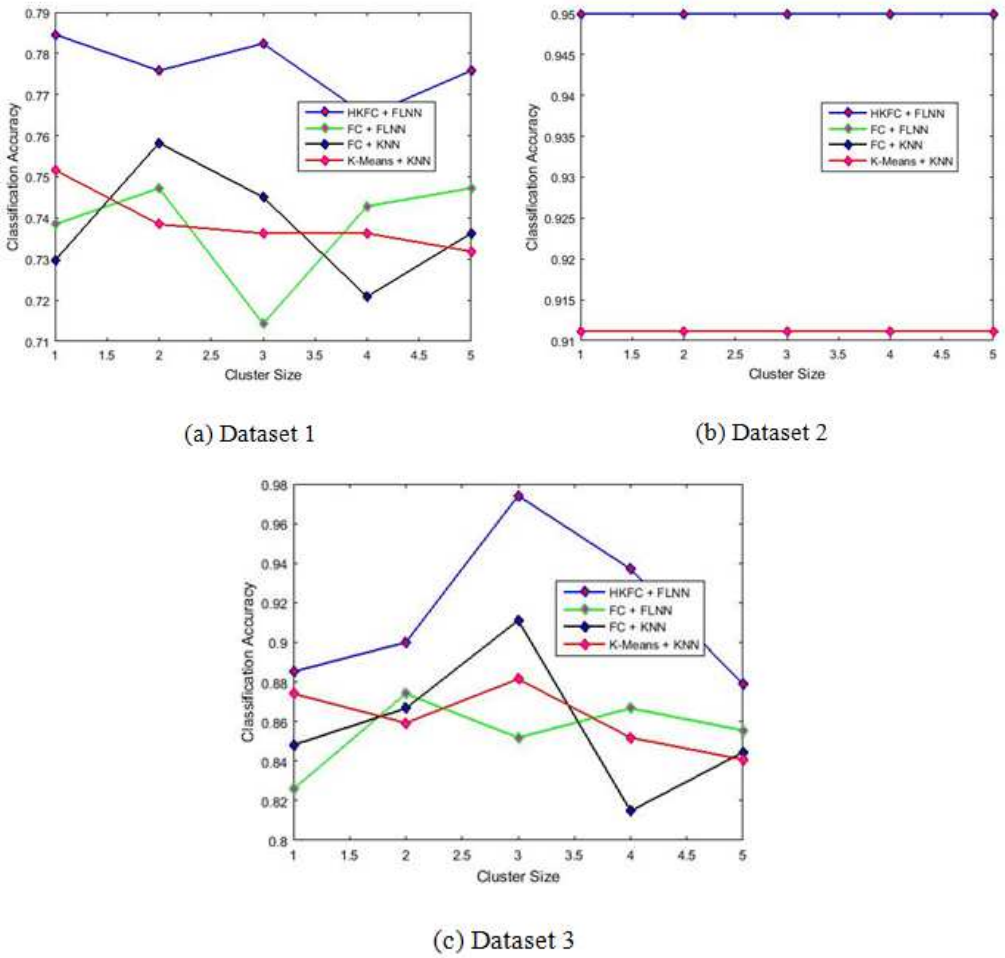

Figure 7: Classification Accuracy Vs Cluster size

DISCUSSIONS

This section summarizes the overall comparative results based on the performance obtained in section 4.4 by providing the best metrics obtained in all the comparative techniques, as given in table 2.

Table 2: Performance Comparison

		HKFC+FLNN	*FC+FLNN*	*FC+KNN*	*K-Means+KNN*
Dataset 1	*MSE*	**0.0344**	0.0903	0.0998	0.0535
	Classification accuracy	**0.8022**	0.7473	0.7582	0.7582
Dataset 2	*MSE*	**0.0048**	0.0057	0.0101	0.0061
	Classification accuracy	**0.95**	0.9111	0.9111	0.9111
Dataset 3	*MSE*	**0.2754**	0.3362	0.3007	0.3079
	Classification accuracy	**0.9741**	0.8741	0.9111	0.9074

As presented in table 2, the best results obtained for the analysis made with different percentages of missing data and cluster sizes; in all the comparative techniques using the three datasets is listed. For the dataset 1, when the existing methodologies, FC+FLNN, FC+KNN and K-Means+KNN had MSE of 0.0903, 0.0998 and 0.0535, the proposed technique, HKFC+FLNN had MSE of just 0.0344. For the same dataset, the classification accuracy was 0.8022 in

HKFC+FLNN, while, FC+FLNN, FC+KNN and K-Means+KNN had only 0.7473, 0.7582 and 0.7582, respectively. For datasets 2 and 3, when HKFC+FLNN could produce MSE of 0.0048, 0.2754 and accuracy of 0.95, 0.9741, the existing FC+FLNN could offer MSE of 0.0057, 0.3362 and accuracy of 0.9111, 0.8741, for the same datasets. As a result of comparison, it is observed that the proposed approach had MSE of 0.0344, 0.0048, 0.2754, and classification accuracy of 0.8022, 0.95 and 0.9741, for dataset 1, dataset 2, and dataset 3, respectively. Hence, it can be concluded that the proposed HKFC+FLNN approach had better performance than the compared existing techniques.

CONCLUSIONS

In this paper, a technique for imputing missing data and classification is presented using a hybrid prediction approach, designed by adopting MKFC and a novel FLNN model. FLNN is designed by the integration of LOA in the weight update equation of a feed forward network that employs LM training algorithm, to improve the search space. Considering the missing data as the class attributes, HKFC and FLNN model impute the missing elements. Once the missing attribute values are obtained, classification is done using FLNN, which effectively assigns the neuron weights for the accurate classification. Three datasets, such as heart disease, iris and wine, from the UCI machine learning repository are considered for the experimentation and the results are compared with three existing approaches, FC+FLNN, FC+KNN and K-Means+KNN using the parameters, MSE and accuracy. From the comparative analysis, the proposed HKFC+FLNN technique could attain MSE of 0.0344, 0.0048, 0.2754, and classification accuracy of 0.8022, 0.95 and 0.9741, for the datasets, heart disease, iris, and wine, respectively. Thus, it can be proved that HKFC+FLNN are effective in imputing missing data and classification than the existing techniques.

REFERENCES

1. *Zhixu Li, Lu Qin, Hong Cheng, Xiangliang Zhang, and Xiaofang Zhou, "TRIP: An Interactive Retrieving-Inferring Data Imputation Approach", IEEE Transaction on Knowledge and Data Engineering, vol. 27, no. 9, pp. 2550-2563, August 2015.*
2. *Jose Luis, Sancho-Gomez, Anibal R.Figueiras Vidal and Michel Verleysen, "K nearest neighbours with mutual information for simultaneous classification and missing data imputation", Neuro computing, vol. 72, no. 9, pp. 1483-1493, 2009.*
3. *Chandan Gautam and Vadlamani Ravi, "Data imputation via evolutionary computation, clustering and a neural network", Neurocomputing, vol. 156, pp. 134-142, May 2015.*
4. *Loris Nanni, Alessandra Lumini and Sheryl Brahnam, "A classifier ensemble approach for the missing feature problem", Artificial Intelligence in Medicine, vol. 55, no. 1, pp. 37-50, May 2012.*
5. *Mr.Papendra Kumar And Mr. Suresh Kumar, Analyze The Medical Image By, Fuzzy Clustering Algorithms Through Edge Detection Methods, International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR), Volume 1, Issue 2, November - December 2011, pp. 1-8*
6. *Laura Folguera, Jure Zupan, Daniel Cicerone and Jorge F. Magallanes, "Self-organizing maps for imputation of missing data in incomplete data matrices", Chemometrics and Intelligent Laboratory Systems, vol. 143, pp. 146-151, 2015.*
7. *Jaemun Sima, Ohbyung Kwon and Kun Chang Lee, "Adaptive pairing of classifier and imputation methods based on the characteristics of missing values in datasets", Expert Systems with Applications, vol. 46, pp. 485–493, 2016.*
8. *Gerhard Tutz a,∗, Shahla Ramzan, "Improved methods for the imputation of missing data by nearest neighbor methods", Computational Statistics and Data Analysis, vol. 90, pp. 84-99, October 2015.*
9. *Md. Geaur Rahman and Md Zahidul Islam, "Missing value imputation using a fuzzy clustering based EM approach",*

Knowledge and Information Systems, vol. 46, no. 2, pp. 389-422, February 2016.

10. *Jing Tian, Bing Yu, Dan Yu and Shilong Ma, "Missing data analyses: a hybrid multiple imputation algorithm using Gray System Theory and entropy based on clustering", Applied Intelligence, vol. 40, no. 2, pp. 376-388, March 2014.*
11. *Marcilio CP de Souto, Pablo A Jaskowiak and Ivan G Costa, "Impact of missing data imputation methods on gene expression clustering and classification", BMC Bioinformatics, vol. 16, February 2015.*
12. *Xiaofeng Zhu, Shichao Zhang, Zili Zhang, and Zhuoming Xu, "Missing Value Estimation for Mixed-Attribute Data Sets", IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 1, pp. 110-121, January 2011.*
13. *Estevam R. Hruschka Jr, Eduardo R. Hruschka and Nelson F. F. Ebecken, "Bayesian networks for imputation in classification problems", Journal of Intelligent Information Systems, vol. 29, no. 3, pp. 231-252, December 2007.*
14. *Ruey-Ling Yeh, Ching Liu, Ben-Chang Shia, Yu-Ting Cheng and Ya-Fang Huwang, "Imputing manufacturing material in data mining", Journal of Intelligent Manufacturing, vol. 19, no. 1, pp. 109-118, February 2008.*
15. *Tae Yeon Kwon and Yousung Park, "A new multiple imputation method for bounded missing values" Statistics & Probability Letters, vol. 107, pp. 204-209, December 2015.*
16. *Jianhua Wu, Qinbao Song and Junyi Shen, "An Novel Association Rule Mining Based Missing Nominal Data Imputation Method", In proceedings of IEEE International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel or Distributed Computing", pp. 244-249, August 2007.*
17. *R. Devi Priya and R. Sivaraj, "Imputation of Discrete and Continuous Missing Values in Large Datasets Using Bayesian Based Ant Colony Optimization", Arabian Journal for Science and Engineering, vol. 41, no. 12, pp. 4981–4993, December 2016.*
18. *Jose M. Jerez, Ignacio Molina and Pedro J. Garcia-Laencina, "Missing data imputation using statistical and machine learning methods in a real breast cancer problem", Artificial Intelligence in Medicine, vol. 50, no. 2, pp. 105-115, October 2010.*
19. *Ibrahim Berkan Aydilek and Ahmet Arslan, "A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm", Information Sciences, vol. 233, pp. 25–35, 2013.*
20. *Pilsung Kang, "Locally linear reconstruction based missing value imputation for supervised learning", Neurocomputing, vol. 118, pp. 65-78, October 2013.*
21. *Ingunn Myrtveit, Erik Stensrud and Ulf H. Olsson, "Analyzing Data Sets with Missing Data: An Empirical Evaluation of Imputation Methods and Likelihood-Based Methods", IEEE Transactions on Software Engineering, vol. 27, no. 11, pp. 999-1013, August 2002.*
22. *Yongsong Qin, Shichao Zhang, Xiaofeng Zhu, Jilian Zhang and Chengqi Zhang, "Semi-parametric optimization for missing data imputation", Applied Intelligence, vol. 27, no. 1, pp. 79-88, August 2007.*
23. *Fabio Lobato, Claudomiro Sales, Igor Araujo, Vincent Tadaiesky, Lilian Dias, Leonardo Ramos and Adamo Santana, "Multi-Objective Genetic Algorithm For Missing Data Imputation", Pattern recognition letters, vol. 68, pp. 126-131, December 2015.*
24. *Bankat M. Patil, Ramesh C. Joshi, Durga Toshniwal, "Missing Value Imputation Based on K-Mean Clustering with Weighted Distance", Contemporary Computing, vol. 94, pp. 600-609, 2010.*
25. *Julian Luengo, Jose A. Saez and Francisco Herrera, "Missing data imputation for fuzzy rule-based classification systems", Soft computing, vol.16, no. 5, pp. 863-881, 2012.*

26. *C. J. Carmona, J. Luengo and P. Gonzalez and M. J. del Jesus, "A Preliminary Study on Missing Data Imputation in Evolutionary Fuzzy Systems of Subgroup Discovery", In proceedings of IEEE International Conference on Fuzzy Systems, pp. 1-7, June 2012.*
27. *Hsin-Chien Huang, Yung-Yu Chuang,Chu-Song Chen, "Multiple Kernel Fuzzy Clustering", IEEE Transactions on Fuzzy Systems, Vol. 20, No. 1, pp. 120-134, February 2012.*
28. *Maziar Yazdani, Fariborz Jolai, "Lion Optimization Algorithm (LOA): A nature-inspired metaheuristic algorithm", Journal of Computational Design and Engineering, Vol. 3, no. 1, pp. 24–36, January 2016.*
29. *Satish Chander, P. Vijaya, Praveen Dhyani, "MKF-Firefly:Hybridization of Firefly and Multiple Kernel-Based Fuzzy C-Means Algorithm",International Journal of Advanced Research in Computer & Communication Engineering, vol. 5, no. 7, pp. 213-216, 2016.*
30. *Manoj Kumar and Charul Bhatnagar, "Crowd Behavior Recognition Using Hybrid Tracking Model and Genetic algorithm Enabled Neural Network", International Journal of Computational Intelligence Systems, Vol. 10, no. 1, pp. 234–246, 2017.*
31. *P Vijaya, G Raju, SK Ray, "Artificial neural network-based merging score for Meta search engine", Journal of Central South University, vol. 23, no. 10, pp. 2604-2615, 2016.*